(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 111312275 A
(43)申请公布日 2020.06.19

(21)申请号 202010090988.0

(22)申请日 2020.02.13

(71)申请人 大连理工大学
地址 116024 辽宁省大连市高新园区凌工路2号

(72)发明人 王鹤 陈喆 殷福亮

(74)专利代理机构 大连东方专利代理有限责任公司 21212
代理人 姜玉蓉 李洪福

(51)Int.Cl.
G10L 21/0272(2013.01)
G10L 21/0308(2013.01)
G10L 21/0216(2013.01)
G10L 25/24(2013.01)
G10L 17/00(2013.01)
G10L 17/02(2013.01)
G10L 17/04(2013.01)
G10L 25/78(2013.01)
G10L 25/84(2013.01)

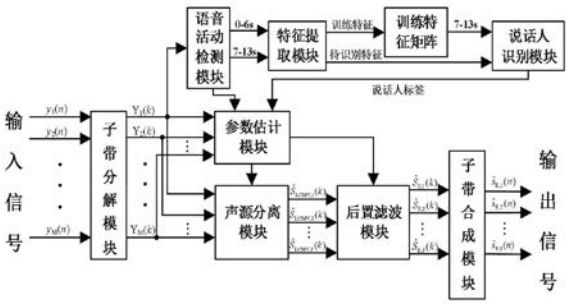
权利要求书3页 说明书12页 附图4页

(54)发明名称

一种基于子带分解的在线声源分离增强系统

(57)摘要

本发明公开了一种基于子带分解的在线声源分离增强系统，具体包括子带分解模块、语音活动检测模块、特征提取模块、说话人识别模块、参数估计模块、声源分离模块、后置滤波模块和子带合成模块。该系统利用识别出的说话人单独发声的片段估计对应声源的相对传递函数RTF，实现了实时的相对传递函数RTF估计，同时降低了其他声源信号对某个特定声源相对传递函数RTF估计的干扰；同时该系统提高了传统KNN说话人识别的准确率，并且在噪声干扰较大时也能有较高的识别准确率。



CN 111312275 A

1.一种基于子带分解的在线声源分离增强系统,其特征在于包括:

子带分解模块,对麦克风接收到的信号进行分帧和缓存处理得到缓存信号,对缓存信号进行子带分解得到子带信号;

语音活动检测模块,接收子带分解模块传送的子带信号,利用子带信号估计临界频带信噪比,对所有临界频带上信噪比求和得到当前帧信号的总信噪比,如果总信噪比大于信噪比阈值则判断当前帧的子带信号为语音并输出,否则,判断当前帧的子带信号为噪声并更新噪声的临界频带能量同时输出噪声子带信号;

特征提取模块,接收语音活动检测模块输出的语音子带信号、提取该子带信号的梅尔倒谱系数,先提取训练时间段的子带信号的特征作为训练特征,在识别阶段提取待识别子带信号的特征作为待识别特征;

说话人识别模块,在识别阶段利用K最近邻算法将待识别特征与训练特征比较得到语音子带信号的说话人标签;

参数估计模块,接收语音活动检测模块输出的语音子带信号、噪声子带信号以及说话人识别模块传送的说话人标签,估计噪声子带信号的噪声功率谱矩阵,在识别阶段读取说话人标签信息并根据语音子带信号估计出该说话人的相对传递函数;

声源分离模块,接收子带分解模块传送的子带信号、参数估计模块传送的相对传递函数矩阵和噪声功率谱矩阵,采用线性约束最小方差(LCMV)算法获取LCMV滤波系数矩阵,将LCMV滤波系数矩阵作用于输入子带信号得到分离后各个声源的子带信号;

后置滤波模块,接收声源分离模块传送的子带信号以及参数估计模块传送的相对传递函数矩阵和噪声功率谱矩阵,利用相对传递函数矩阵和噪声功率谱矩阵估计残留噪声功率谱矩阵和目标信号功率谱矩阵,采用多说话人维纳后置滤波(MWPF)算法获取后置滤波系数矩阵,将后置滤波系数矩阵作用于声源分离模块输出的子带信号得到最终的子带信号;

子带合成模块,接收后置滤波模块传送的子带信号、对该子带信号进行缓存处理得到子带缓存信号,对子带缓存信号进行子带合成得到各个声源的时域信号。

2.根据权利要求1所述的一种基于子带分解的在线声源分离增强系统,其特征还在于:所述子带分解模块对麦克风接收到的信号以一定的采样频率采样后得到$y_i(n)$,i=1,2,...,M,M是麦克风的数目,对该信号进行分帧、缓存得到$y_i'(l,n)$,缓存的长度为N,则子带分解后的信号为

$$Y_i(l,k)=\sum_{s=0}^{2D-1}\sum_{r=0}^{5}y_i'(N-(r\cdot 2D+s))\cdot h(r\cdot 2D+s)\cdot e^{j\frac{2\pi ks}{2D}},\ k=0,1,...,2D-1 \qquad (1)$$

其中,l表示帧号,k表示子带,D是子带数目的一半,在本发明中设为D=160,N=6×2D,h(n)为分析滤波器的系数

$$h(n)=\frac{2}{\pi}\cdot\frac{(0.54-0.46\cos\frac{2n+1}{N})\cdot\sin(\pi\cdot\frac{2n-N+1}{4D})}{2n-N+1},\ n=0,1,\ldots,N-1 \qquad (2)$$

其中式(1)采用如下算法计算:

$$temp(s)=\sum_{r=0}^{5}y_i'(N-(r\cdot 2D+s))\cdot h(r\cdot 2D+s),\ s=0,1,...,2D-1 \qquad (3)$$

$$Y_i(l,k)=\sum_{s=0}^{2D-1} temp(s)\cdot e^{j\frac{2\pi ks}{2D}} \tag{4}$$

其中，式(4)采用快速傅里叶变换实现，在计算子带信号时，只需计算前面一半即可，后面一半根据共轭对称性直接得出，即

$$Y_i(l,k)=\begin{cases} Y_i(l,k) & k=0,1,...,D \\ Y_i(l,2D-k) & k=D+1,...,2D-1 \end{cases} 。 \tag{5}$$

3.根据权利要求1所述的一种基于子带分解的在线声源分离增强系统，其特征还在于：所述说话人识别模块获取子带信号的说话人标签采用如下方式：如果当前帧信号检测为语音，则用该帧语音信号计算出的梅尔倒谱系数(MFCC)特征v(l)与训练特征矩阵T中每一行的前12维特征计算欧式距离

$$d_i(l)=\sqrt{\sum_{n=1}^{12}|v(l,n)-T(i,n)|^2},\ i=1,2,...,L \tag{15}$$

对所有的$d_i(l)$排序，找出最小的K个并记录其标号为index(k)，k＝1,2,...,K，则当前帧信号经过KNN判决的说话人标签为

$$C(l)\leftarrow \arg\max_{c\in S_c}\sum_{k=1}^{K}\frac{1}{d_{index(k)}(l)}\delta(c,T(index(k),13)) \tag{16}$$

对式(16)的解释为：在最小的K个距离中，求对应标签相同的距离的倒数之和，和最大的标签被判定为当前帧的说话人标签，其中，S_c＝{1,2,...,J}为所有说话人标签的集合，T(index(k),13)表示第k个最小的距离所对应的说话人标签，δ函数定义为

$$\delta(a,b)=\begin{cases} 1 & if\ a=b \\ 0 & if\ a\neq b \end{cases} \tag{17}$$

4.根据权利要求1所述的一种基于子带分解的在线声源分离增强系统，其特征还在于：所述参数估计模块估计说话人的相对传递函数RTF具体采用如下方式：若当前帧检测为语音信号并且识别为第j个说话人，则利用该帧子带信号估计第j个声源的相对传递函数，其估计方法如下：计算各路子带信号和第一个麦克风子带信号之间的互功率谱密度

$$\Phi_{y_1y_i,j}(l,k)=\alpha_2\cdot\Phi_{y_1y_i,j}(l-1,k)+(1-\alpha_2)\cdot Y_1^*(l,k)Y_i(l,k),\ i=1,2,...,M \tag{18}$$

其中，α_2为帧间的平滑系数，取值在0～1之间，则第j个声源和各个麦克风之间的相对传递函数为

$$\tilde{A}_{i,j}(l,k)=\frac{\overline{\Phi_{y_1y_1,j}(k)\Phi_{y_1y_i,j}(k)}-\overline{\Phi_{y_1y_1,j}(k)}\,\overline{\Phi_{y_1y_i,j}(k)}}{\overline{\Phi^2_{y_1y_1,j}(k)}-\overline{\Phi_{y_1y_i,j}(k)}^2},\ i=1,2,...,M \tag{19}$$

式中，符号上面的横线代表帧间取平均，即

$$\overline{Z(k)}=\frac{1}{L'}\sum_{l=1}^{L'}Z(l,k) \tag{20}$$

其中，L'是取平均的帧数，用各个声源的相对传递函数构成阶数为M×J的RTF矩阵

$$\tilde{\mathbf{A}}=\begin{bmatrix}\tilde{A}_{1,1} & \cdots & \tilde{A}_{1,J}\\ \vdots & \ddots & \vdots\\ \tilde{A}_{M,1} & \cdots & \tilde{A}_{M,J}\end{bmatrix} \tag{21}$$

其中，J是声源的数目。

估计噪声子带信号的噪声功率谱矩阵NPSD采用如下方式：若当前帧麦克风接收信号检测为噪声帧，则用该帧子带信号估计NPSD矩阵，估计方法为

$$\Phi_v(l,k)=\gamma\cdot\Phi_v(l-1,k)+(1-\gamma)\cdot y(l,k)y^H(l,k) \tag{22}$$

其中，$y(l,k)=[Y_1(l,k),\ldots,Y_M(l,k)]^T$是输入信号向量，$[\,]^T$表示矩阵的转置，$[\,]^H$表示矩阵的共轭转置，$\gamma$为帧间平滑系数，取值在0～1之间。

5.根据权利要求4所述的一种基于子带分解的在线声源分离增强系统，其特征还在于：所述声源分离模块利用线性约束最小方差(LCMV)准则计算滤波系数，对麦克风接收的子带信号进行滤波得到分离后各个声源的子带信号，根据多说话人LCMV准则，最优滤波系数矩阵为

$$\mathbf{W}_{\mathrm{LCMV}}=\boldsymbol{\Phi}_v^{-1}\tilde{\mathbf{A}}(\tilde{\mathbf{A}}^H\boldsymbol{\Phi}_v^{-1}\tilde{\mathbf{A}})^{-1} \tag{23}$$

其中，$[\,]^{-1}$表示矩阵求逆，在式(23)中，为了保证矩阵求逆的顺利进行，需要满足以下两个条件：(a)NPSD矩阵必须满秩；(b)RTF矩阵的列秩必须为J。认为各个麦克风接收到的噪声信号不相干，则满足条件(a)，此外，麦克风数目多于声源数目，并且认为各个声源的传递函数线性无关，故可满足条件(b)，J为声源个数。

将最优滤波系数矩阵作用于麦克风接收到的各路子带信号，得到分离后各个声源的子带信号

$$\begin{aligned}\hat{\mathbf{s}}_{\mathrm{LCMV}}&=\begin{bmatrix}\hat{S}_{\mathrm{LCMV},1} & \hat{S}_{\mathrm{LCMV},2} & \cdots & \hat{S}_{\mathrm{LCMV},J}\end{bmatrix}\\&=\mathbf{W}_{\mathrm{LCMV}}^{\mathrm{H}}\mathbf{y}\\&=\mathbf{s}_{\mathrm{E}}+\mathbf{v}_{\mathrm{R}}\end{aligned} \tag{24}$$

其中，分离后的信号包含两部分，一部分是各个声源的目标语音信号s_E，另一部分是残留的噪声信号v_R。

一种基于子带分解的在线声源分离增强系统

技术领域

[0001] 本发明涉及语音信号处理技术领域，尤其涉及一种基于子带分解的在线声源分离增强系统。

背景技术

[0002] 语音交流是人类生活中必不可少的一部分，语音表达的信息比文字更加直接。近年来，智能手机、智能音箱等可以进行人机交互的智能设备得到了广泛应用，这些设备可以识别人们发出的交互指令，方便了人们的生活。但是，当有多个人(一般为2～4个)同时讲话时，因为语音间的相互干扰，导致智能设备的语音识别率明显降低，因此，需要将多个声源同时发出的语音分离出来，智能设备才能对特定声源发出的语音进行识别。

[0003] Markovich等在文献[1]中提出一种可以抑制多个语音干扰源的语音增强方法，该方法采用广义旁瓣消除(GSC)架构实现，如图1所示，分为三部分：固定波束形成器(FBF)、阻塞矩阵(BM)和自适应噪声消除器(ANC)，FBF将信号延迟求和得到初步增强的单路信号，BM利用目标声源和干扰声源的声传递函数实现，能使干扰信号和噪声源通过，阻止目标声源信号通过，ANC采用自适应的方法进一步抑制干扰和噪声信号。但是文献[1]的缺陷是使用一个GSC波束形成器只能增强单个特定声源的信号，抑制其他方向的干扰和噪声。若要同时分离出多个不同声源的信号，需要用多个不同的波束形成器，计算量较大。

[0004] Schwartz等在文献[2]提出一种基于最小均方误差(MMSE)准则的多声源分离方法，通过求解使各个声源的期望信号与实际分离出信号之间均方误差最小的约束优化问题得到各个频带上的最优滤波器，该滤波器可分解成一个多声源GSC波束形成器和一个后置滤波器，对麦克风阵列接收到的信号进行多声源波束形成和后置滤波得到分离后的各个声源信号。其中，波束形成器利用各个声源的声传递函数导出。其中文献[2]是的缺陷是估计声传递函数需要利用各个声源单独发声的语音片段，因此，在估计之前要人工手动标记出各个声源单独发声的片段，无法实时处理，进而不能实现在线声源分离。

[0005] 因此传统的基于波束形成的声源分离方法在估计某个特定声源的参数时，需要提前标记出各个声源单独发声的语音片段，无法实现实时的参数估计和声源分离。另一方面，传统波束形成方法只能增强某一声源方向的信号，同时抑制其它方向的干扰和噪声信号，不能同时分离出多个声源的信号。

发明内容

[0006] 针对上述问题，本发明提出一种基于子带分解的在线声源分离增强系统，该系统利用说话人识别技术识别出各个声源单独发声的片段，然后实时估计出各个声源对于所有麦克风的相对传递函数(RTF)，利用多说话人线性约束最小方差(LCMV)方法同时分离出各个声源的语音信号，最后采用多声源维纳后置滤波(MWPF)方法抑制各个声源语音信号中的残留噪声，提高分离出的各个语音信号的信干噪比(SINR)。该系统具体包括：

[0007] 子带分解模块，对麦克风接收到的信号进行分帧和缓存处理得到缓存信号，对缓

存信号进行子带分解得到子带信号；

[0008] 语音活动检测模块，接收子带分解模块传送的子带信号，利用子带信号估计临界频带信噪比，对所有临界频带上信噪比求和得到当前帧信号的总信噪比，如果总信噪比大于信噪比阈值则判断当前帧的子带信号为语音并输出，否则，判断当前帧的子带信号为噪声并更新噪声的临界频带能量同时输出噪声子带信号；

[0009] 特征提取模块，接收语音活动检测模块输出的语音子带信号、提取该子带信号的梅尔倒谱系数，先提取训练时间段的子带信号的特征作为训练特征，在识别阶段提取待识别子带信号的特征作为待识别特征；

[0010] 说话人识别模块，在识别阶段利用K最近邻算法将待识别特征与训练特征比较得到语音子带信号的说话人标签；

[0011] 参数估计模块，接收语音活动检测模块输出的语音子带信号、噪声子带信号以及说话人识别模块传送的说话人标签，估计噪声子带信号的噪声功率谱矩阵，在识别阶段读取说话人标签信息并根据语音子带信号估计出该说话人的相对传递函数；

[0012] 声源分离模块，接收子带分解模块传送的子带信号、参数估计模块传送的相对传递函数矩阵和噪声功率谱矩阵，采用LCMV算法获取LCMV滤波系数矩阵，将LCMV滤波系数矩阵作用于输入子带信号得到分离后各个声源的子带信号；

[0013] 后置滤波模块，接收声源分离模块传送的子带信号以及参数估计模块传送的相对传递函数矩阵和噪声功率谱矩阵，利用相对传递函数矩阵和噪声功率谱矩阵估计残留噪声功率谱矩阵和目标信号功率谱矩阵，采用MWPF算法获取后置滤波系数矩阵，将后置滤波系数矩阵作用于声源分离模块输出的子带信号得到最终的子带信号；

[0014] 子带合成模块，接收后置滤波模块传送的子带信号、对该子带信号进行缓存处理得到子带缓存信号，对子带缓存信号进行子带合成得到各个声源的时域信号。

[0015] 进一步的，所述子带分解模块对麦克风接收到的信号以一定的采样频率采样后得到$y_i(n)$，i＝1,2,...,M,M是麦克风的数目，对该信号进行分帧、缓存得到$y_i'(l,n)$，缓存的长度为N，则子带分解后的信号为

[0016] $$Y_i(l,k)=\sum_{s=0}^{2D-1}\sum_{r=0}^{5}y_i'(N-(r\cdot 2D+s))\cdot h(r\cdot 2D+s)\cdot e^{j\frac{2\pi ks}{2D}},\ k=0,1,...,2D-1 \qquad (1)$$

[0017] 其中，l表示帧号，k表示子带，D是子带数目的一半，在本发明中设为D＝160，N＝6×2D，h(n)为分析滤波器的系数

[0018] $$h(n)=\frac{2}{\pi}\cdot\frac{(0.54-0.46\cos\frac{2n+1}{N})\cdot\sin(\pi\cdot\frac{2n-N+1}{4D})}{2n-N+1},\ n=0,1,\ ...,N-1 \qquad (2)$$

[0019] 其中式(1)采用如下算法计算：

[0020] $$temp(s)=\sum_{r=0}^{5}y_i'(N-(r\cdot 2D+s))\cdot h(r\cdot 2D+s),\ s=0,1,...,2D-1 \qquad (3)$$

[0021] $$Y_i(l,k)=\sum_{s=0}^{2D-1}temp(s)\cdot e^{j\frac{2\pi ks}{2D}} \qquad (4)$$

[0022] 其中，式(4)采用快速傅里叶变换实现，在计算子带信号时，只需计算前面一半即可，后面一半根据共轭对称性直接得出，即

[0023] $$Y_i(l,k)=\begin{cases}Y_i(l,k) & k=0,1,...,D\\ Y_i(l,2D-k) & k=D+1,...,2D-1\end{cases}。\qquad(5)$$

[0024] 进一步的，所述说话人识别模块获取子带信号的说话人标签采用如下方式：如果当前帧信号检测为语音，则用该帧语音信号计算出的梅尔倒谱系数(MFCC)特征v(l)与训练特征矩阵T中每一行的前12维特征计算欧式距离

[0025] $$d_i(l)=\sqrt{\sum_{n=1}^{12}|v(l,n)-T(i,n)|^2},\ i=1,2,...,L\qquad(15)$$

[0026] 对所有的$d_i(l)$排序，找出最小的K个并记录其标号为index(k)，k＝1,2,...,K，则当前帧信号经过KNN判决的说话人标签为

[0027] $$C(l)\leftarrow\arg\max_{c\in S_c}\sum_{k=1}^{K}\frac{1}{d_{index(k)}(l)}\delta(c,T(index(k),13))\qquad(16)$$

[0028] 对式(16)的解释为：在最小的K个距离中，求对应标签相同的距离的倒数之和，和最大的标签被判定为当前帧的说话人标签，其中，S_c＝{1,2,...,J}为所有说话人标签的集合，T(index(k),13)表示第k个最小的距离所对应的说话人标签，δ函数定义为

[0029] $$\delta(a,b)=\begin{cases}1 & if\ a=b\\ 0 & if\ a\neq b\end{cases}\qquad(17)$$

[0030] 进一步的，所述参数估计模块估计说话人的相对传递函数RTF具体采用如下方式：若当前帧检测为语音信号并且识别为第j个说话人，则利用该帧子带信号估计第j个声源的相对传递函数，其估计方法如下：计算各路子带信号和第一个麦克风子带信号之间的互功率谱密度

[0031] $\Phi_{y1yi,j}(l,k)=\alpha_2\cdot\Phi_{y1yi,j}(l-1,k)+(1-\alpha_2)\cdot Y_1^*(l,k)Y_i(l,k)$，i＝1,2,...,M (18)

[0032] 其中，α_2为帧间的平滑系数，取值在0～1之间，则第j个声源和各个麦克风之间的相对传递函数为

[0033] $$\tilde{A}_{i,j}(l,k)=\frac{\overline{\Phi_{y_1y_1,j}(k)\Phi_{y_1y_i,j}(k)}-\overline{\Phi_{y_1y_1,j}(k)\Phi_{y_1y_i,j}(k)}}{\overline{\Phi^2_{y_1y_1,j}(k)}-\overline{\Phi_{y_1y_i,j}(k)}^2},\ i=1,2,...,M\qquad(19)$$

[0034] 式中，符号上面的横线代表帧间取平均，即

[0035] $$\overline{Z(k)}=\frac{1}{L'}\sum_{l=1}^{L'}Z(l,k)\qquad(20)$$

[0036] 其中，L'是取平均的帧数，用各个声源的相对传递函数构成阶数为M×J的RTF矩阵

[0037] $$\tilde{\mathbf{A}}=\begin{bmatrix}\tilde{A}_{1,1} & \cdots & \tilde{A}_{1,J}\\ \vdots & \ddots & \vdots\\ \tilde{A}_{M,1} & \cdots & \tilde{A}_{M,J}\end{bmatrix}\qquad(21)$$

[0038] 其中，J是声源的数目。

[0039] 估计噪声子带信号的噪声功率谱矩阵NPSD采用如下方式：若当前帧麦克风接收信号检测为噪声帧，则用该帧子带信号估计NPSD矩阵，估计方法为

[0040] $\Phi_v(l,k)=\gamma\cdot\Phi_v(l-1,k)+(1-\gamma)\cdot y(l,k)y^H(l,k)$ (22)

[0041] 其中，$y(l,k)=[Y_1(l,k),...,Y_M(l,k)]^T$是输入信号向量，$[\,]^T$表示矩阵的转置，$[\,]^H$表示矩阵的共轭转置，$\gamma$为帧间平滑系数，取值在0～1之间。

[0042] 进一步的，所述声源分离模块利用LCMV准则计算滤波系数，对麦克风接收的子带信号进行滤波得到分离后各个声源的子带信号，根据多说话人LCMV准则，最优滤波系数矩阵为

[0043] $$\mathbf{W}_{\mathrm{LCMV}}=\boldsymbol{\Phi}_{\mathrm{v}}^{-1}\tilde{\mathbf{A}}(\tilde{\mathbf{A}}^{\mathrm{H}}\boldsymbol{\Phi}_{\mathrm{v}}^{-1}\tilde{\mathbf{A}})^{-1} \tag{23}$$

[0044] 其中，$[\,]^{-1}$表示矩阵求逆，在式(23)中，为了保证矩阵求逆的顺利进行，需要满足以下两个条件：(a)NPSD矩阵必须满秩；(b)RTF矩阵的列秩必须为J。认为各个麦克风接收的到噪声信号不相干，则满足条件(a)，此外麦克风数目多于声源数目，并且认为各个声源的传递函数线性无关，故可满足条件(b)，J为声源个数。

[0045] 将最优滤波系数矩阵作用于麦克风接收到的各路子带信号，得出分离后各个声源的子带信号

[0046] $$\begin{aligned}\hat{\mathbf{s}}_{\mathrm{LCMV}}&=\left[\hat{S}_{\mathrm{LCMV},1}\quad\hat{S}_{\mathrm{LCMV},2}\quad\cdots\quad\hat{S}_{\mathrm{LCMV},J}\right]\\&=\mathbf{W}_{\mathrm{LCMV}}^{\mathrm{H}}\mathbf{y}\\&=\mathbf{s}_{\mathrm{E}}+\mathbf{v}_{\mathrm{R}}\end{aligned} \tag{24}$$

[0047] 其中，分离后的信号包含两部分，一部分是各个声源的目标语音信号s_E，另一部分是残留的噪声信号v_R。

[0048] 由于采用了上述技术方案，本发明提供的一种基于子带分解的在线声源分离增强系统，该系统利用识别出的说话人单独发声的片段估计对应声源的相对传递函数RTF，实现了实时的相对传递函数RTF估计，同时降低了其他声源信号对某个特定声源相对传递函数RTF估计的干扰；同时该系统提高了传统KNN说话人识别的准确率，并且在噪声干扰较大时也能有较高的识别准确率。

附图说明

[0049] 为了更清楚地说明本申请实施例或现有技术中的技术方案，下面将对实施例或现有技术描述中所需要使用的附图作简单地介绍，显而易见地，下面描述中的附图仅仅是本申请中记载的一些实施例，对于本领域普通技术人员来讲，在不付出创造性劳动的前提下，还可以根据这些附图获得其他的附图。

[0050] 图1为本发明背景技术中GSC结构图；

[0051] 图2为本发明中系统的结构原理图；

[0052] 图3为本发明中联合判决流程图；

[0053] 图4为本发明中声源位置图；

[0054] 图5为本发明中分离前第一个麦克风的语音波形图；

[0055] 图6为本发明中分离后第一个说话人的语音波形图；

[0056] 图7为本发明中分离前第一个麦克风的语谱图；

[0057] 图8为本发明中分离后第一个说话人的语谱图。

具体实施方式

[0058] 为使本发明的技术方案和优点更加清楚，下面结合本发明实施例中的附图，对本发明实施例中的技术方案进行清楚完整的描述：

[0059] 如图2所示的一种基于子带分解的在线声源分离增强系统，包括子带分解模块、语音活动检测模块、特征提取模块、说话人识别模块、参数估计模块、声源分离模块、后置滤波模块和子带合成模块。其中，参数估计模块包括对相对传递函数RTF和噪声功率谱矩阵NPSD的估计。首先，对所有麦克风采集到的各路信号进行子带分解，得到子带信号，然后将第一个麦克风的子带信号通入语音活动检测模块，判断当前帧子带信号是否为语音。把前6秒定义为训练时间段，将前6秒信号分为J段，每段信号只有单个声源发出声音，用第一个麦克风前6秒信号中检测为语音的信号帧提取各个说话人的梅尔倒谱系数(MFCC)特征构成训练特征矩阵，将7～13秒定义为识别阶段，在7～13秒的信号中，对每一帧检测为语音的信号提取MFCC特征进行说话人识别，将识别结果和各路麦克风的子带信号通入参数估计模块估计出各个声源的RTF。在只有噪声期间估计NPSD，然后得到LCMV滤波系数，对各路输入信号滤波得到分离后的各个声源信号，最后，用计算出的MWPF滤波系数对分离后的信号滤波并经过子带合成得到各个声源的输出信号。这里要求前13秒只能有单个声源轮流发出声音，13秒之后各个声源可以同时发出声音。

[0060] 进一步的，子带分解模块的工作原理是：对麦克风接收到的信号以16kHz的采样频率采样后得到$y_i(n)$，i＝1,2,...,M，M是麦克风的数目，对该信号进行分帧(本发明中帧长设为160)、缓存得到$y_i'(l,n)$，缓存的长度为N，则子带分解后的信号为

[0061] $$Y_i(l,k)=\sum_{s=0}^{2D-1}\sum_{r=0}^{5}y_i'(N-(r\cdot 2D+s))\cdot h(r\cdot 2D+s)\cdot e^{j\frac{2\pi ks}{2D}},\ k=0,1,...,2D-1 \qquad (1)$$

[0062] 其中，l表示帧号，k表示子带，D是子带数目的一半，在本发明中设为D＝160，N＝6×2D，h(n)为分析滤波器的系数

[0063] $$h(n)=\frac{2}{\pi}\cdot\frac{(0.54-0.46\cos\frac{2n+1}{N})\cdot\sin(\pi\cdot\frac{2n-N+1}{4D})}{2n-N+1},\ n=0,1,\ ...,N-1 \qquad (2)$$

[0064] 在本发明中，式(1)的实现分为以下两步：

[0065] $$temp(s)=\sum_{r=0}^{5}y_i'(N-(r\cdot 2D+s))\cdot h(r\cdot 2D+s),\ s=0,1,...,2D-1 \qquad (3)$$

[0066] $$Y_i(l,k)=\sum_{s=0}^{2D-1}temp(s)\cdot e^{j\frac{2\pi ks}{2D}} \qquad (4)$$

[0067] 其中，式(4)可用快速傅里叶变换实现。此外，在计算子带信号时，只需计算前面一半即可，后面一半可以根据共轭对称性直接得出，即

[0068] $$Y_i(l,k)=\begin{cases}Y_i(l,k) & k=0,1,...,D\\ Y_i(l,2D-k) & k=D+1,...,2D-1\end{cases} \qquad (5)$$

[0069] 进一步的，语音活动检测模块的工作原理是：将对应频率为0.3～4kHz的子带分成16个临界频带，各个临界频带的起始子带如表1所示。第一个麦克风的子带信号在各个临界频带上的平均能量为

[0070] $$E_p(l,i)=\alpha_1\cdot E_p(l-1,i)+\frac{1-\alpha_1}{b(i)-a(i)+1}\sum_{k=a(i)}^{b(i)}\left|Y_1(l,k)\right|^2,\ i=1,2,...,16 \qquad (6)$$

[0071] 其中，a(i)、b(i)分别是第i个临界频带的起始子带点，α_1是帧间的平滑系数，取值在0～1之间，在本发明中设为$\alpha_1=0.9$。

[0072] 在本发明中，用前6帧信号初始化噪声的临界频带能量$E_n(1,i)$，令其与输入信号的临界频带能量相等。则各个临界频带上的信噪比为

[0073] $$snr(l,i)=\left\lfloor\left(10\log_{10}\frac{E_p(l,i)}{E_n(l,i)}\right)\Big/0.375+0.5\right\rfloor \qquad (7)$$

[0074] 其中，$\lfloor\ \rfloor$表示向下取整。

[0075] 对所有临界频带上信噪比求和得到当前帧信号的总信噪比，在本发明中，设信噪比的阈值为30，若总信噪比大于信噪比阈值则判断当前帧信号为语音信号，否则，判断当前帧信号为噪声信号并更新噪声的临界频带能量，更新公式为

[0076] $E_n(1,i)=\beta_1\cdot E_n(1-1,i)+(1-\beta_1)\cdot E_p(1,i)$ (8)

[0077] 其中，β_1是帧间的平滑系数，在本发明中设为$\beta_1=0.9$。

[0078] 表1临界频带的起始子带点

[0079]

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
开始	2	4	6	9	12	15	18	21	25	29	34	39	45	52	61	70
结束	3	5	8	11	14	17	20	24	28	33	38	44	51	60	69	79

[0080] 进一步的，特征提取模块的工作原理是：提取的特征为梅尔倒谱系数(MFCC)，该特征在梅尔频率尺度上划分，近似人类的听觉系统，广泛应用于语音识别和说话人识别。频率和梅尔频率的转换关系为

[0081] $$\begin{cases}F_{Mel}(f)=2595\lg(1+f/700)\\ F_{Mel}^{-1}(b)=700(e^{\frac{b}{1125}}-1)\end{cases} \qquad (9)$$

[0082] 在本发明中，利用第一个麦克风检测为语音帧的子带信号$Y_1(1,k)$计算梅尔倒谱系数，首先计算子带信号在前D个子带的能量

[0083] $E(1,k)=Y_1^*(1,k)*Y_1(1,k)$，$k=0,1,...,D-1$ (10)

[0084] 然后，计算梅尔滤波器组的频率响应

[0085] $$H_r(k)=\begin{cases}0, & k<f(r-1)\\ \dfrac{k-f(r-1)}{f(r)-f(r-1)}, & f(r-1)\le k\le f(r)\\ \dfrac{f(r+1)-k}{f(r+1)}, & f(r)\le k\le f(r+1)\\ 0, & k>f(r+1)\end{cases},\ r=1,2,...,R \qquad (11)$$

[0086] 其中，R是梅尔滤波器的个数，本发明中设置为R＝40，f(r)是梅尔滤波器的中心频率，表示为

[0087] $$f(r)=\frac{2D}{f_s}F_{Mel}^{-1}(F_{Mel}(f_l)+r\frac{F_{Mel}(f_h)-F_{Mel}(f_l)}{R+1}),\ r=0,1,...,R+1 \quad (12)$$

[0088] 式中，f_s为信号的采样频率，本发明设置为f_s＝16kHz，f_l、f_h分别是梅尔滤波器组可通过的最低、最高频率，本发明设置为f_l＝0.3kHz、f_h＝8kHz。

[0089] 将子带能量通过梅尔滤波器组，得到梅尔能量

[0090] $$S(l,r)=\sum_{k=0}^{D-1}E(l,k)H_r(k),\ r=0,1,...,R-1 \quad (13)$$

[0091] 最后，对梅尔能量取对数后再经过离散余弦变换得到该帧信号的梅尔倒谱系数

[0092] $$\mathrm{mfcc}(l,n)=\sqrt{\frac{2}{R}}\sum_{r=0}^{R-1}\log\left[S(l,r)\right]\cos\left[\frac{\pi n(2r-1)}{2R}\right],\ n=0,1,...,R-1 \quad (14)$$

[0093] 在本发明中，取第2-13维MFCC作为说话人特征向量，维度为12。本发明的特征提取分为两个阶段：训练特征提取、待识别特征提取阶段。前6秒是训练特征提取阶段，将前6秒的信号分为J段，代表了J个声源，每一段信号中包括一个声源单独发声的语音，第j段信号中检测为语音帧的说话人标签为j，j＝1,2,...,J，将每一帧语音信号提取的12维MFCC再加上1维说话人标签作为训练特征矩阵T的一行，所以T的维度为L×13，L是提取MFCC训练特征的总帧数。7-13秒是待识别特征提取阶段，每一帧检测为语音的信号提取12维MFCC后通入说话人识别模块进行说话人识别，得到该帧信号的说话人标签。

[0094] 进一步的，说话人识别模块的工作原理是：在说话人识别阶段，若当前帧信号检测为噪声，则将该帧信号的说话人标签置0，若当前帧信号检测为语音，则用该帧语音信号计算出的MFCC特征v(l)与训练特征矩阵T中每一行的前12维特征计算欧式距离

[0095] $$d_i(l)=\sqrt{\sum_{n=1}^{12}|v(l,n)-T(i,n)|^2},\ i=1,2,...,L \quad (15)$$

[0096] 对所有的$d_i(l)$排序，找出最小的K个并且记录其标号为index(k)，k＝1,2,...,K，则当前帧信号经过KNN判决的说话人标签为

[0097] $$C(l)\leftarrow\arg\max_{c\in S_c}\sum_{k=1}^{K}\frac{1}{d_{index(k)}(l)}\delta(c,T(index(k),13)) \quad (16)$$

[0098] 对式(16)的解释为：在最小的K个距离中，求对应标签相同的距离的倒数之和，和最大的标签被判定为当前帧的说话人标签。其中，S_c＝{1,2,...,J}为所有说话人标签的集合，T(index(k),13)表示第k个最小的距离所对应的说话人标签，δ函数定义为

[0099] $$\delta(a,b)=\begin{cases}1 & if\ a=b\\ 0 & if\ a\neq b\end{cases} \quad (17)$$

[0100] 为了提高说话人识别的准确率，本发明使用多帧联合判决的方法，具体过程如图3所示：首先缓存连续100帧信号的说话人标签，若前50帧中有超过40帧的标签都相同，则将前50帧的标签都设置为该标签，否则，认为前50帧的标签无效，全部置0，同理，对后50帧做相同的处理。若前后50帧的标签都有效且相同，则前50帧的标签保持不变，否则，将前50帧的标签置0。当前100帧判决完成之后，输出前50帧的标签到参数估计模块，然后将后50帧的标签作为前50帧的标签，重新缓存50帧的标签作为后50帧的标签，按照前述判决方法继续

判决，直至说话人识别结束。该方法提高了基于KNN的说话人识别的鲁棒性，在噪声干扰较大时也能有较高的识别率，降低了一帧或几帧的错误判决对后续参数估计的影响。

[0101] 进一步的，参数估计模块的工作原理是：估计声源分离所需要的参数，包括相对传递函数RTF和噪声功率谱密度NPSD。RTF的估计利用每个说话人单独说话的语音片段实现，若当前帧检测为语音信号并且识别为第j个说话人，则利用该帧子带信号估计第j个声源的相对传递函数，估计方法如下：

[0102] 首先，计算各路子带信号和第一个麦克风子带信号之间的互功率谱密度

[0103] $\Phi_{y1yi,j}(l,k)=\alpha_2\cdot\Phi_{y1yi,j}(l-1,k)+(1-\alpha_2)\cdot Y_1^*(l,k)Y_i(l,k),i=1,2,...,M$ (18)

[0104] 其中，α_2为帧间的平滑系数，取值在0～1之间，本专利设为$\alpha_2=0.95$。

[0105] 则第j个声源和各个麦克风之间的相对传递函数为

[0106] $$\tilde{A}_{i,j}(l,k)=\frac{\overline{\Phi_{y_1y_1,j}(k)\Phi_{y_1y_i,j}(k)}-\overline{\Phi_{y_1y_1,j}(k)}\,\overline{\Phi_{y_1y_i,j}(k)}}{\overline{\Phi^2_{y_1y_1,j}(k)}-\overline{\Phi_{y_1y_i,j}(k)}^2},\ i=1,2,...,M \qquad (19)$$

[0107] 式中，符号上面的横线代表帧间取平均，即

[0108] $$\overline{Z(k)}=\frac{1}{L'}\sum_{l=1}^{L'}Z(l,k) \qquad (20)$$

[0109] 其中，L'是取平均的帧数，在本发明设为L'=20。

[0110] 最后，用各个声源的相对传递函数构成阶数为M×J的RTF矩阵

[0111] $$\tilde{\mathbf{A}}=\begin{bmatrix}\tilde{A}_{1,1} & \cdots & \tilde{A}_{1,J}\\ \vdots & \ddots & \vdots\\ \tilde{A}_{M,1} & \cdots & \tilde{A}_{M,J}\end{bmatrix} \qquad (21)$$

[0112] NPSD的估计利用只有噪声的片段实现，若当前帧麦克风接收信号检测为噪声帧，则用该帧子带信号估计NPSD矩阵，估计方法为

[0113] $\Phi_v(l,k)=\gamma\cdot\Phi_v(l-1,k)+(1-\gamma)\cdot y(l,k)y^H(l,k)$ (22)

[0114] 其中，$y(l,k)=[Y_1(l,k),...,Y_M(l,k)]^T$是输入信号向量，$[\,]^T$表示矩阵的转置，$[\,]^H$表示矩阵的共轭转置，$\gamma$为帧间平滑系数，取值在0～1之间，在本发明中取值为$\gamma=0.95$。

[0115] 进一步的，声源分离模块的工作原理是：利用LCMV准则计算滤波系数，然后对麦克风接收的子带信号进行滤波得到分离后各个声源的子带信号。根据多说话人LCMV准则，最优滤波系数矩阵为

[0116] $$\mathbf{W}_{LCMV}=\mathbf{\Phi}_v^{-1}\tilde{\mathbf{A}}(\tilde{\mathbf{A}}^H\mathbf{\Phi}_v^{-1}\tilde{\mathbf{A}})^{-1} \qquad (23)$$

[0117] 其中，$[\,]^{-1}$表示矩阵求逆。在式(23)中，为了保证矩阵求逆的顺利进行，需要满足以下两个条件：(a)NPSD矩阵必须满秩；(b)RTF矩阵的列秩必须为J。在本发明中，认为各个麦克风接收的到噪声信号不相干，可满足条件(a)，此外，本发明中麦克风数目多于声源数目，并且认为各个声源的传递函数线性无关，故可满足条件(b)。

[0118] 最优矩阵每一列的作用是增强该列对应的声源方向的信号，抑制其它声源方向的信号和噪声信号。将最优矩阵作用于麦克风接收到的各路子带信号，可以得出分离后各个

声源的子带信号

$$\begin{aligned}\hat{\mathbf{s}}_{\mathrm{LCMV}} &= \left[\hat{S}_{\mathrm{LCMV},1} \quad \hat{S}_{\mathrm{LCMV},2} \quad \cdots \quad \hat{S}_{\mathrm{LCMV},J}\right] \\ &= \mathbf{W}_{\mathrm{LCMV}}^{\mathrm{H}}\mathbf{y} \\ &= \mathbf{s}_{\mathrm{E}} + \mathbf{v}_{\mathrm{R}}\end{aligned} \tag{24}$$

[0119]

[0120] 其中，分离后的信号包含两部分，一部分是各个声源的目标语音信号s_E，另一部分是残留的噪声信号v_R。

[0121] 进一步的，后置滤波模块的工作原理是：经过LCMV模块分离出的信号中仍然含有部分残留噪声信号，该模块的作用是采用多说话人维纳后置滤波的方法抑制残留噪声信号，进一步提高语音质量。

[0122] 残留噪声的PSD矩阵为

$$\begin{aligned}\mathbf{\Phi}_{\mathrm{VR}} &= \mathbf{W}_{\mathrm{LCMV}}^{\mathrm{H}}\mathbf{\Phi}_{\mathrm{v}}\mathbf{W}_{\mathrm{LCMV}} \\ &= (\hat{\mathbf{A}}^{\mathrm{H}}\mathbf{\Phi}_{\mathrm{v}}^{-1}\hat{\mathbf{A}})^{-1}\end{aligned} \tag{25}$$

[0123]

[0124] 在本发明中，认为各个声源之间是相互独立的，因此各个声源目标语音信号的PSD矩阵可以等效为一个对角矩阵，即

[0125] $\Phi_{SE}=\mathrm{diag}\{\Phi_{SE,1}\Phi_{SE,2}\ldots\Phi_{SE,J}\}$ (26)

[0126] 其中，$\Phi_{SE,j}$可以采用决策方向法估计得出

[0127] $$\phi_{\mathrm{SE},j} = \beta_2\left|\hat{S}_{\mathrm{E},j}(l-1)\right|^2 + (1-\beta_2)\max\left\{\left|\hat{S}_{\mathrm{LCMV},j}(l)\right|^2 - \phi_{\mathrm{VR},j}, 0\right\} \tag{27}$$

[0128] 式中，$\Phi_{VR,j}$是残留噪声PSD矩阵对角线上的第j个元素，β_2是帧间的平滑因子，取值为0～1之间，在本发明中的取值为$\beta_2=0.99$，max(a,b)表示求a、b两者中的最大值。

[0129] 根据最小均方误差准则(MMSE)得到使残留噪声最小的多说话人后置维纳滤波系数矩阵，维度为J×J，表示为

[0130] $W_{WPF}=(\Phi_{SE}+\Phi_{VR})^{-1}\Phi_{SE}$ (28)

[0131] 将滤波系数矩阵作用于LCMV模块的输出信号得到最终的输出信号

[0132] $$\hat{\mathbf{s}}_{\mathrm{E}} = \mathbf{W}_{\mathrm{WPF}}^{\mathrm{H}}\hat{\mathbf{s}}_{\mathrm{LCMV}} \tag{29}$$

[0133] 进一步的，子带合成模块的工作原理是将子带信号合成为时域信号。子带合成的具体操作和子带分解的步骤正好相反，先升采样再进行滤波，最终数据相加实现信号的重构。

[0134] 首先，对子带信号做类似于式(4)的计算得到临时信号

[0135] $$temp_j(i) = \sum_{k=0}^{D-1}\hat{S}_{\mathrm{E},j}(l,k)e^{j\frac{2\pi ki}{2D}},\ i=0,1,\ldots,2D-1 \tag{30}$$

[0136] 用该临时信号更新子带合成缓存信号的缓存区，表示如下

[0137] $$buffer_j(i) = \begin{cases} buffer_j(i-2D), & i=2D,\ldots,N' \\ temp_j(i), & i=0,1,\ldots,2D-1 \end{cases} \tag{31}$$

[0138] 其中，N'是缓存区的长度，在本发明中设为N'=3840，是子带数目的12倍。然后，对缓存信号进行滤波得到子带合成后的信号

[0139] $$\hat{s}_{\mathrm{E},j}(l,i)=\sum_{k=0}^{N'/D-1}h(kD+i)\cdot buffer_j(k\cdot 2D+(k\,\&1)D+i),\ i=0,1,...,D-1 \qquad (32)$$

[0140] 其中，h是子带合成滤波器的系数，在本发明中，该滤波器的系数与子带分解时的分析滤波器系数相同，如式(2)所示。&表示位与运算，运算结果如下

[0141] $$k\,\&1=\begin{cases}1, & if\ k\text{为奇数}\\ 0, & if\ k\text{为偶数}\end{cases} \qquad (33)$$

[0142] 式(32)实现了对子带信号的升采样、滤波和数据相加，最终得到子带合成之后的时域信号。

[0143] 实施例：

[0144] 为验证本发明方法的有效性，本发明测试了三个声源的识别和分离情况。本发明通过Imgae房间冲击响应模型模拟了一个6×6×3的封闭式房间，混响时间T60为0.1、0.3秒。如图4所示，本发明所使用的麦克风阵列为均匀线阵，中心坐标为(3米，3米，1米)，阵元数目为8，阵元的间距为4厘米，三个声源分别位于阵列的正前方、正左方、正右方，并且距离阵列中心的距离均为两米，噪声源位于声源1、3之间，距离阵列中心的距离也为两米。声源是从TIMIT数据库[3]中随机选取的三个不同说话人，每个说话人选取2段时长为2秒、一段时长为4秒的纯净语音信号，信号的采样频率是16kHz。前6秒时，三个说话人依次说出各自的第一段时长为2秒的语音，停顿1秒，7～13秒时，三个说话人依次说出各自的第二段时长为2秒的语音，停顿1秒，在14-18秒时，三个说话人同时说出各自的时长为4秒的语音。噪声源选取为高斯白噪声，分别测试输入信号信噪比为0dB、10dB、20dB时，说话人识别(7～13秒)的正确率和声源分离(7～18秒)后输出信号的SINR。其中，输出信号的SINR定义为

[0145] $$\mathrm{oSINR}=\frac{1}{L_{总}}\sum_{l=1}^{L_{总}}10\log_{10}\frac{\sum_{k=0}^{D-1}\left\|\mathbf{s}_{\mathrm{E}}(l,k)\right\|^2}{\sum_{k=0}^{D-1}\left\|\hat{\mathbf{s}}_{\mathrm{E}}(l,k)-\mathbf{s}_{\mathrm{E}}(l,k)\right\|^2} \qquad (34)$$

[0146] 其中，$L_{总}$为输入信号的总帧数，帧长设置为160，$\|\cdot\|$表示计算向量的2范数，$\log_{10}()$表示计算以10为底的常用对数。

[0147] 此时，采用本发明提出的声源分离方法对麦克风阵列接收到的信号进行说话人识别、声源分离和噪声抑制。在不同输入信噪比的情况下，说话人识别的正确率如表2所示，分离前后信号的SINR如表3(混响时间0.1秒)、表4(混响时间0.3秒)所示。在信噪比为20dB、混响时间0.1秒时，分离前和分离后第一个说话人的7-18秒语音波形如图5、图6所示，语谱图如图7、图8所示。

[0148] 由表3、表4可见，本发明提出的说话人识别方法在不同信噪比和混响时间时均有较高的正确率。在混响为0.1秒时，本发明提出的方法能够使分离后信号的SINR提升13dB左右，在混响为0.3秒时，本发明提出的方法能够使分离后信号的SINR提升11dB左右。根据上述结果和分离前后的语音波形可知本发明提出的分离方法具有较好的分离能力，分离后的语音中噪声残留较少，并且语音的失真不大。

[0149] 表2说话人识别的正确率

[0150]

SNR / 正确率 / T60	0dB	10dB	20dB
0.1 秒	90.06%	94.68%	97.32%
0.3 秒	88.75%	91.55%	93.47%

[0151] 表3混响时间T60＝0.1秒时的oSINR

[0152]

SNR / oSINR / Method	0dB	10dB	20dB

[0153]

不分离	-12.78dB	-8.65dB	-4.86dB
本发明方法	0.76dB	4.59dB	8.93dB

[0154] 表4混响时间T60＝0.3秒时的oSINR

[0155]

SNR / oSINR / Method	0dB	10dB	20dB
不分离	-12.78dB	-8.65dB	-4.86dB
本发明方法	-0.28dB	3.96dB	7.73dB

[0156] (1)本发明所提出的声源分离方法对麦克风阵列的阵型没有限制,可以用其它形状的阵列(如均匀圆阵、L型阵列等)代替,同样能完成本发明的目的。

[0157] (2)本发明中所提出的说话人识别的部分,可以用其它的说话人识别算法(如i-vector等)代替,同样能完成本发明的目的。

[0158] (3)本发明中所提出的利用LCMV进行语音分离的部分,LCMV结构滤波器可以用GSC结构滤波器结构代替,同样能完成本发明的目的。

[0159] (4)本发明中所提出的多说话人(多通道)维纳后置滤波器,可以用J个单说话人(单通道)维纳后置滤波器代替,同样能完成本发明的目的。

[0160] (5)本发明中所提出的多说话人维纳后置滤波器,可以用其它后置滤波器(如LSA等)代替,同样能完成本发明的目的。

[0161] 以上所述,仅为本发明较佳的具体实施方式,但本发明的保护范围并不局限于此,任何熟悉本技术领域的技术人员在本发明揭露的技术范围内,根据本发明的技术方案及其发明构思加以等同替换或改变,都应涵盖在本发明的保护范围之内。

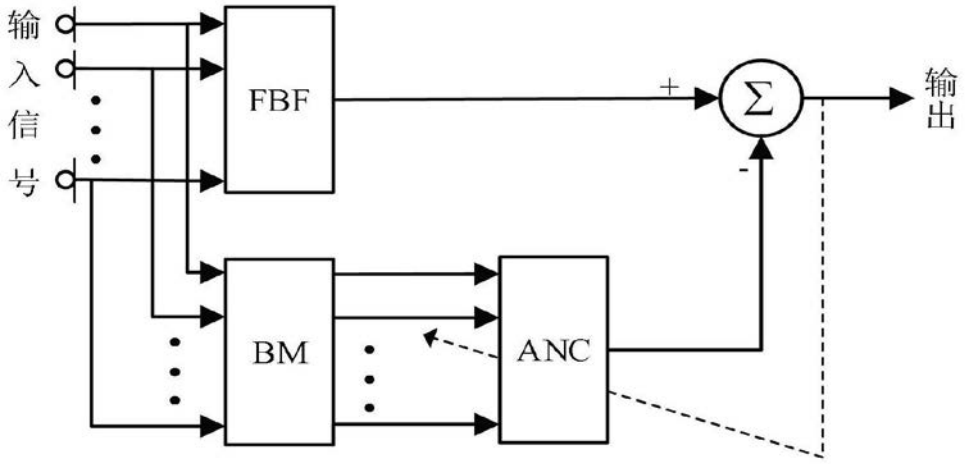

输入信号
FBF
BM
ANC
Σ
+
-
输出

图1

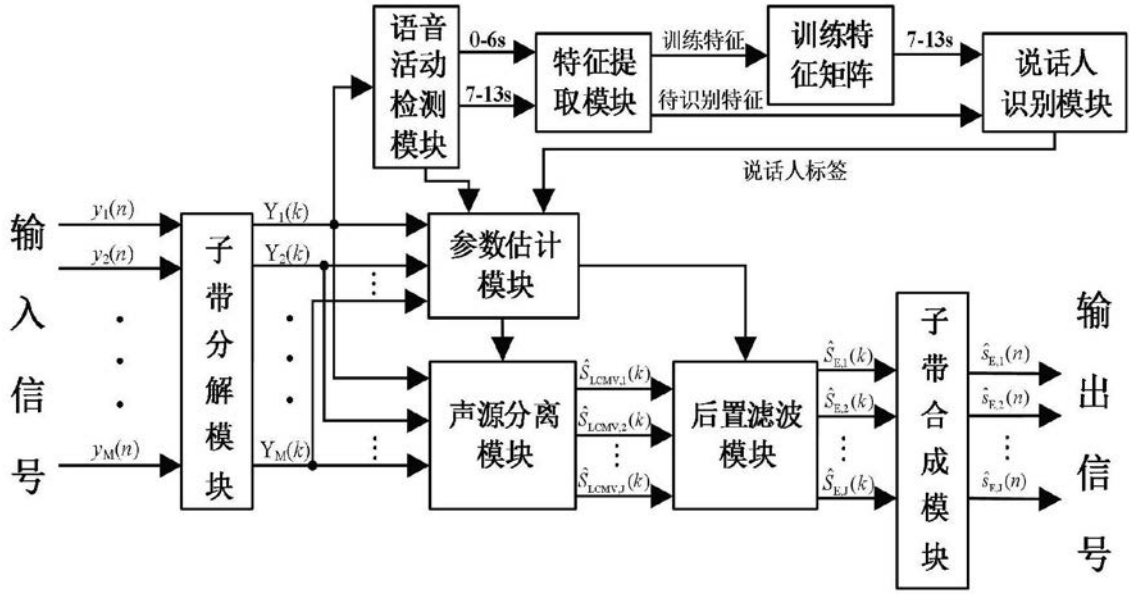

语音活动检测模块
0-6s
7-13s
特征提取模块
训练特征
待识别特征
训练特征矩阵
7-13s
说话人识别模块
说话人标签
输入信号
$y_1(n)$
$y_2(n)$
$y_M(n)$
子带分解模块
$Y_1(k)$
$Y_2(k)$
$Y_M(k)$
参数估计模块
声源分离模块
$\hat{S}_{LCMV,1}(k)$
$\hat{S}_{LCMV,2}(k)$
$\hat{S}_{LCMV,J}(k)$
后置滤波模块
$\hat{S}_{E,1}(k)$
$\hat{S}_{E,2}(k)$
$\hat{S}_{E,J}(k)$
子带合成模块
$\hat{s}_{E,1}(n)$
$\hat{s}_{E,2}(n)$
$\hat{s}_{E,J}(n)$
输出信号

图2

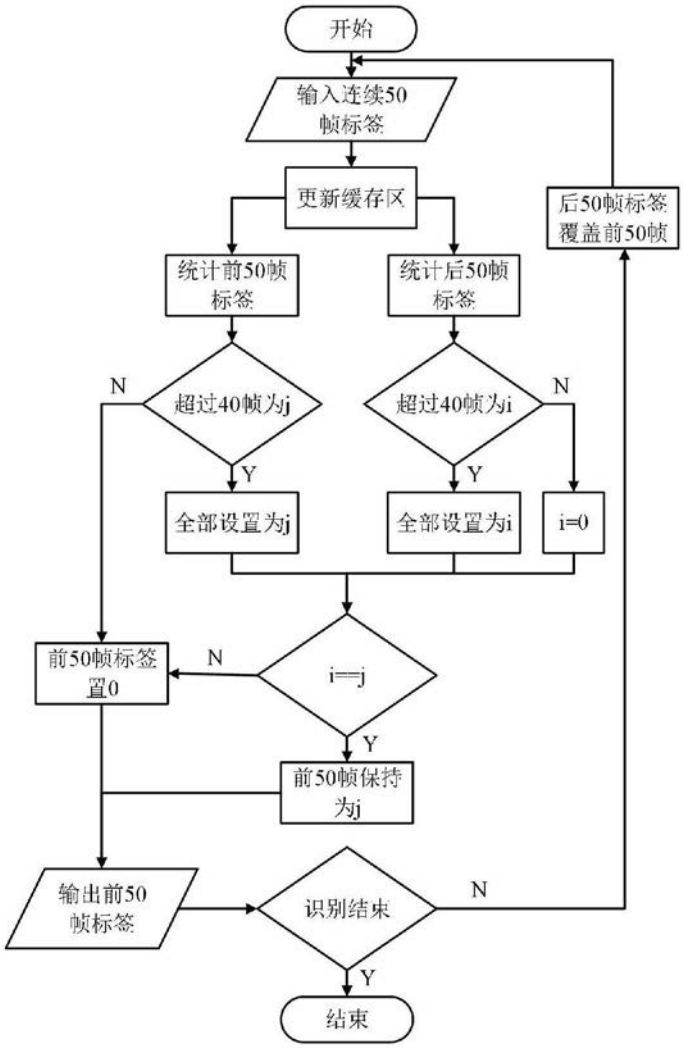

开始
输入连续50帧标签
更新缓存区
后50帧标签覆盖前50帧
统计前50帧标签
统计后50帧标签
超过40帧为j
超过40帧为i
N
Y
全部设置为j
全部设置为i
i=0
前50帧标签置0
i==j
前50帧保持为j
输出前50帧标签
识别结束
结束

图3

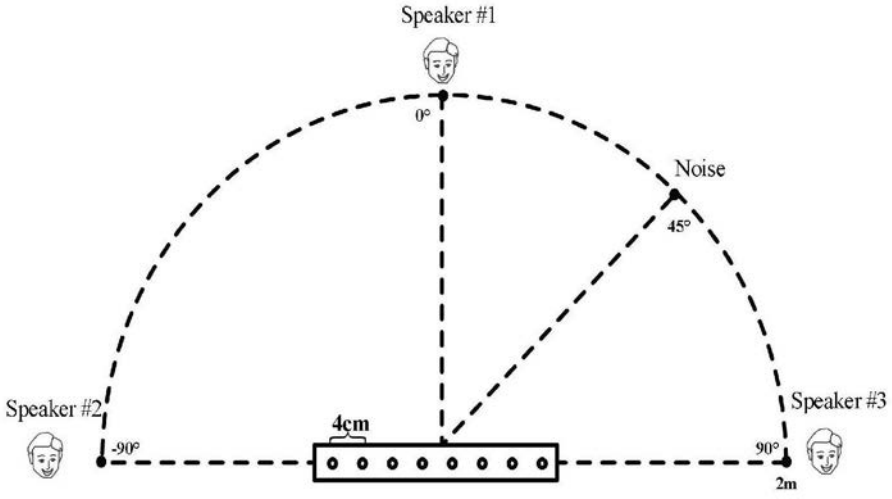

Speaker #1
0°
Noise
45°
Speaker #2
-90°
4cm
90°
Speaker #3
2m

图4

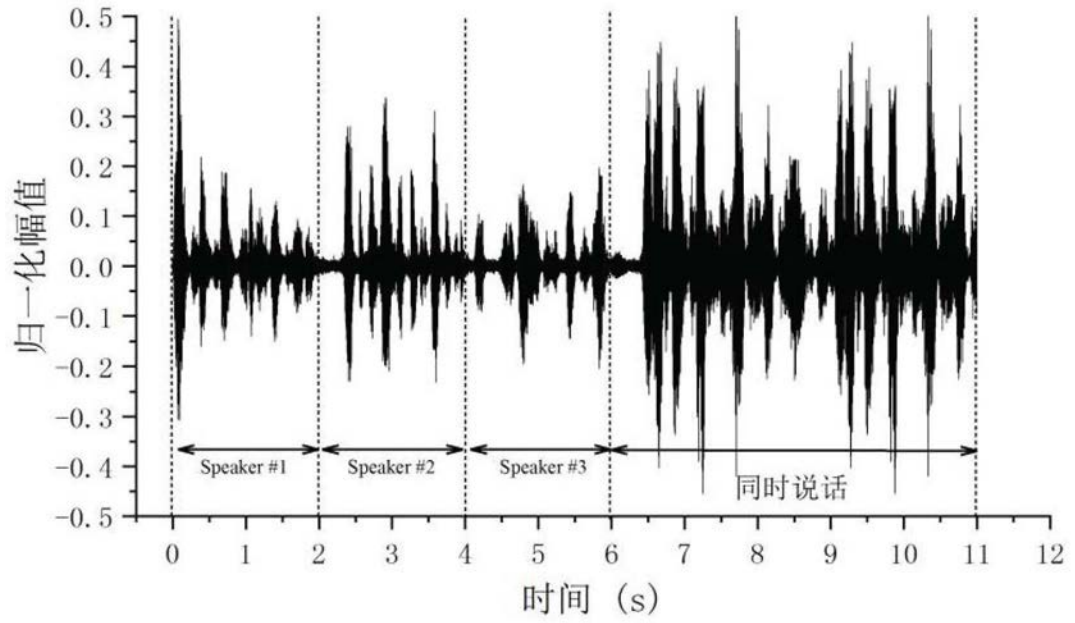

归一化幅值
0.5
0.4
0.3
0.2
0.1
0.0
-0.1
-0.2
-0.3
-0.4
-0.5
Speaker #1
Speaker #2
Speaker #3
同时说话
0 1 2 3 4 5 6 7 8 9 10 11 12
时间（s）

图5

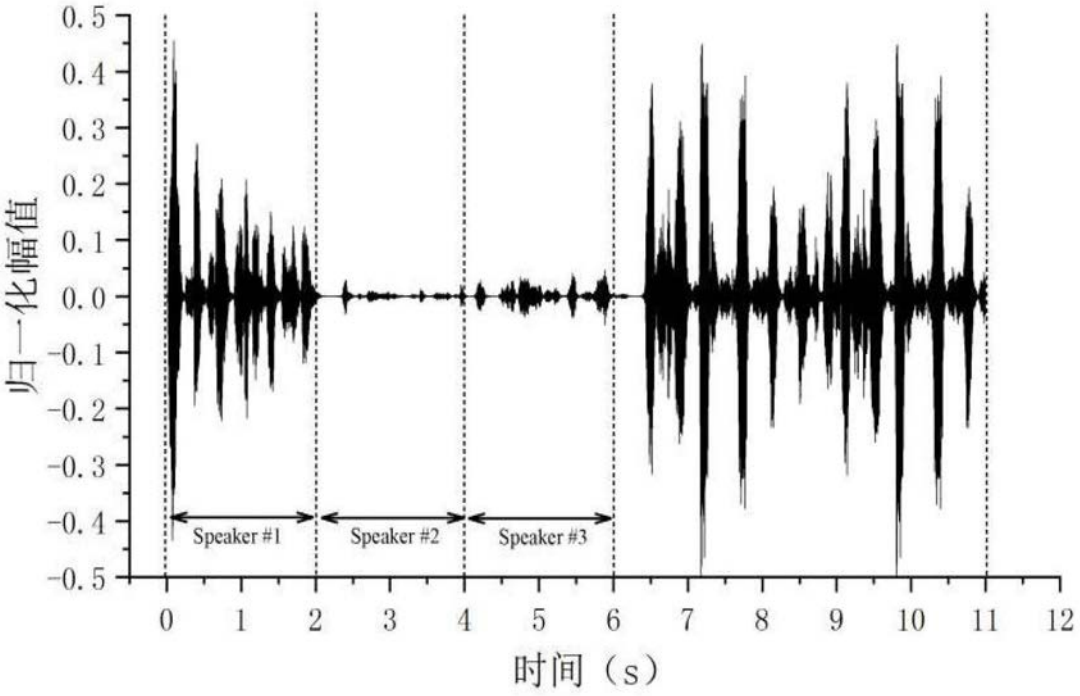

归一化幅值
0.5
0.4
0.3
0.2
0.1
0.0
-0.1
-0.2
-0.3
-0.4
-0.5
Speaker #1
Speaker #2
Speaker #3
0 1 2 3 4 5 6 7 8 9 10 11 12
时间（s）

图6

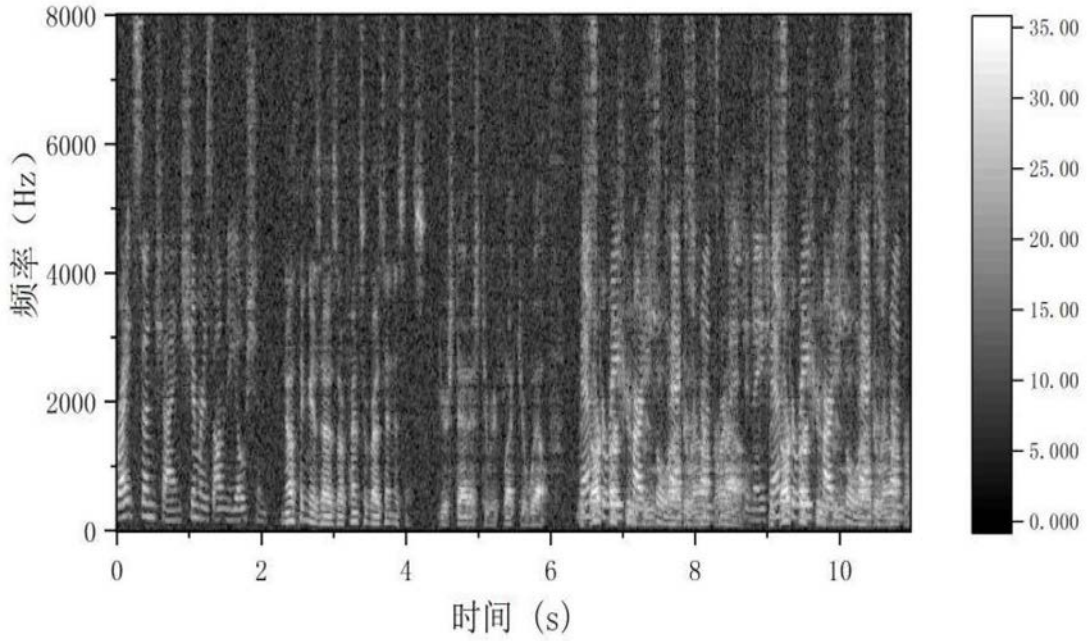

8000
6000
4000
2000
0
频率（Hz）
0
2
4
6
8
10
时间（s）
35.00
30.00
25.00
20.00
15.00
10.00
5.000
0.000

图7

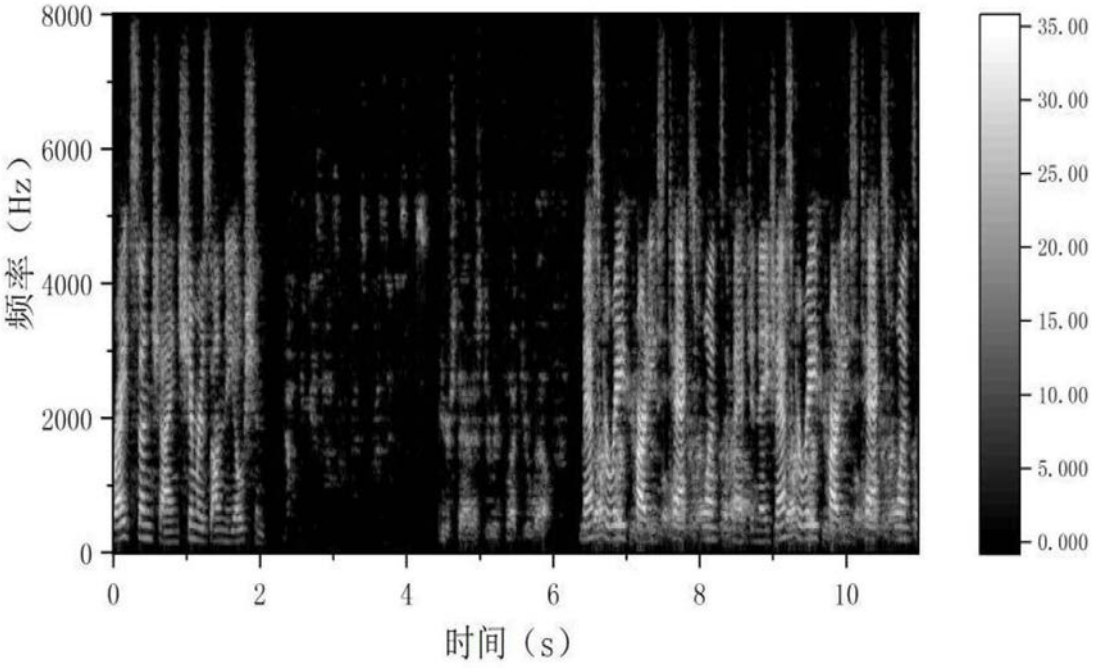

8000
6000
4000
2000
0
频率（Hz）
0
2
4
6
8
10
时间（s）
35.00
30.00
25.00
20.00
15.00
10.00
5.000
0.000

图8